Zerbino *et al. Genome Biology* (2015) 16:56
DOI 10.1186/s13059-015-0621-5

Genome Biology

SOFTWARE Open Access

The Ensembl Regulatory Build

Daniel R Zerbino*, Steven P Wilder, Nathan Johnson, Thomas Juettemann and Paul R Flicek*

Abstract

Most genomic variants associated with phenotypic traits or disease do not fall within gene coding regions, but in regulatory regions, rendering their interpretation difficult. We collected public data on epigenetic marks and transcription factor binding in human cell types and used it to construct an intuitive summary of regulatory regions in the human genome. We verified it against independent assays for sensitivity. The Ensembl Regulatory Build will be progressively enriched when more data is made available. It is freely available on the Ensembl browser, from the Ensembl Regulation MySQL database server and in a dedicated track hub.

Background

Despite our increasing knowledge of genomes and their variants, the downstream effects of sequence variants and the affected cellular mechanisms are still poorly understood. In particular, a large number of the variants identified in genome-wide association studies are located in non-protein coding regions [1], and are presumed to affect gene expression regulation. Similarly, it has been proposed that a significant fraction of the potential for phenotypic adaptation lies within the regulatory elements of the genome [2,3].

There is still much to learn about the dynamic regulation of gene expression [3,4]. *Cis*-regulatory elements are short segments of the genome that either recruit transcription factors (TFs) or affect the properties of the messenger RNA as it is being transcribed [5]. Gene expression is also highly tied to transmissible epigenetic marks [6-8]. The DNA molecule and the histone proteins it is wrapped around can be modified with various additions, such as methyl, acetyl or phosphate groups. These alterations have been shown to provide crucial markers of developmental diseases [9] and cancer [10]. Finally, the three-dimensional conformation of the DNA molecule also affects its activity. In particular, it determines which regions are accessible to outside molecules [11], and which regions are in physical proximity to each other despite being distant in the genomic sequence [12].

Various experimental techniques help us identify the epigenetic markers of the genome and the putative underlying *cis*-regulatory elements. Chromatin immunoprecipitation (ChIP) coupled with either genome-wide tiling microarrays (ChIP-chip [13]) or direct high-throughput sequencing (ChIP-Seq [14-16]) make it possible to perform genome-wide and protein-specific measurements of DNA binding, as well as detect a range of histone modifications. Other methodologies have been developed to identify modified cytosine bases, ranging from array-based approaches such as MeDIP-chip [17], through to more exhaustive approaches such as whole-genome bisulphite sequencing [18]. Regions of open chromatin can be mapped using formaldehyde-assisted isolation of regulatory elements (FAIRE) [19], nuclease digestion by DNase1 coupled with high-throughput sequencing (DNase-seq) [20] or assaying transposase-accessible chromatin (ATAC-seq) [21].

Significant efforts to provide genome-wide maps of histone modifications have already proved successful in elucidating some of the basic patterns associated with promoter and enhancer regions [14,15,22,23]. In addition to an explosion of small and medium-scale studies producing this type of data, large-scale projects like ENCODE [24,25], Roadmap Epigenomics [26], and Blueprint [27] are releasing large amounts of valuable data into the public domain. With the promise of even higher sequencing throughput, genome-wide epigenomic datasets will only become more abundant.

One important challenge is to bring together and standardize these studies, in order to integrate all the information into a coordinated regulatory annotation of the genome. To address this challenge we developed the Ensembl Regulatory Build, within the Ensembl

* Correspondence: zerbino@ebi.ac.uk; flicek@ebi.ac.uk
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

project [28], to provide a high-level overview of the regulatory activity of the genome. Through this process, we annotate putative regulatory regions from public experimental data, and associate these regions with regulatory function.

Results

We defined genomic regions of interest characterised by biochemical activity through a four-step Regulatory Build process that combined all available data, summarised in Figure 1.

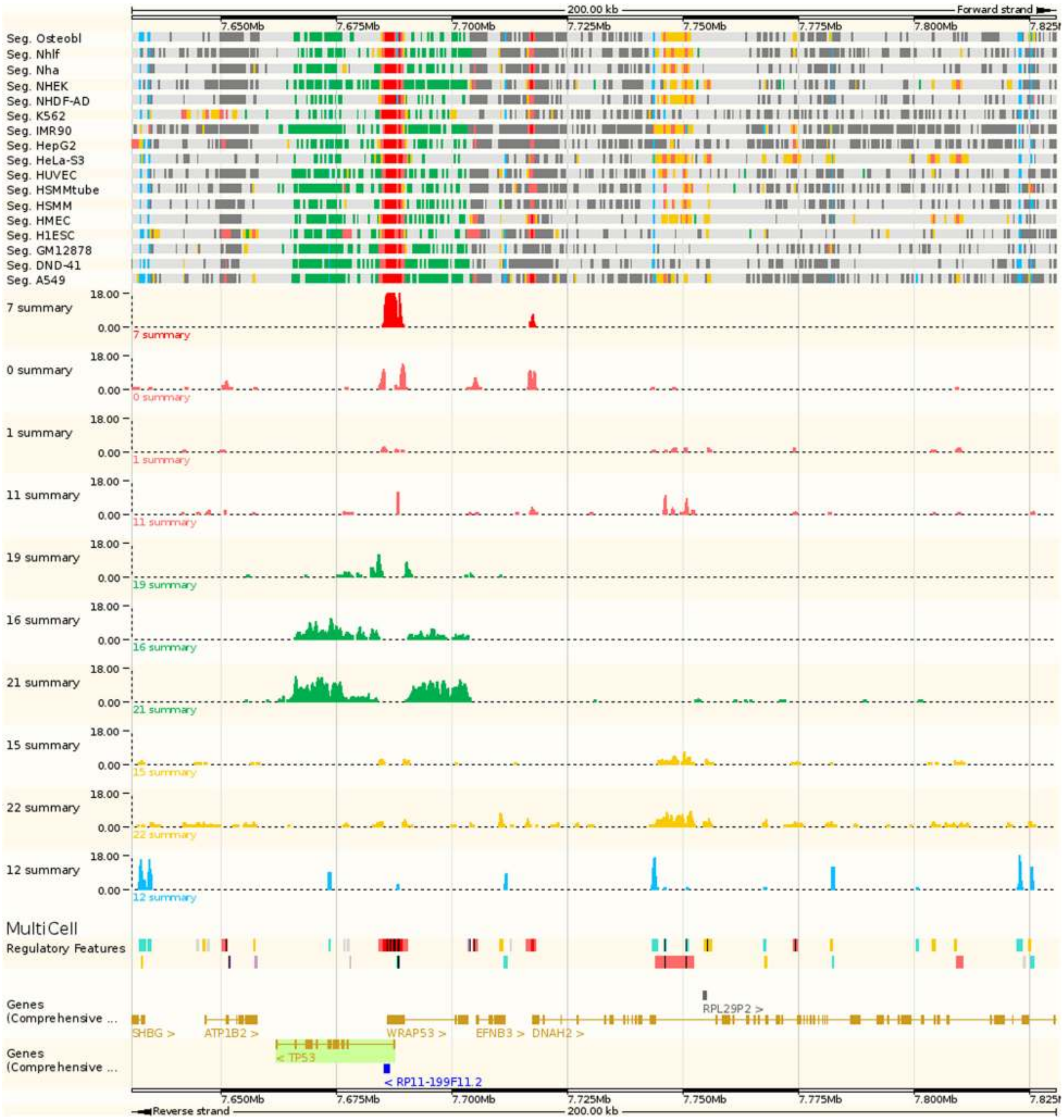

Figure 1 The Regulatory Build process. In a first step we run segmentation software across multiple cell types. For each cell type and at each base pair, the genome is assigned a state, identified by an arbitrary number assigned by the segmentation software. We assign to each state a non-unique functional label, represented by its color on the browser, as shown at the top. For each state at each base pair, we compute the number of cell types sharing that state at that position, as shown in the center of the figure. Having selected relevant states and set some thresholds, we define regions of interest, which are the foundation of the regulatory build. These regions are then complemented with unannotated ChIP-Seq transcription factor binding site peaks and unannotated DNase1 hypersensitivity sites.

We first reduced all the experimental data for each cell type into a cell type-specific annotation of the genome. This can be done with segmentation tools, such as Segway [29] or ChromHMM [30]. In a first training pass, these algorithms take as input a set of genome-wide assays, and detect recurring signal patterns (referred to as 'states'). In a segmentation pass, for each cell type at each base pair of the genome, they determine the most likely underlying state, based on local experimental measurements.

By overlapping these segmentation states, produced by unsupervised machine learning, with known genomic features, we assigned them functional labels, such as 'predicted promoter with TSS' (where TSS is transcription start site), 'predicted transcribed region', 'predicted promoter flank', 'predicted enhancer', 'CTCF enriched', 'predicted repressed', 'predicted low activity', 'predicted heterochromatin'. To ensure the broadest applicability of our approach, we minimized the use of known epigenetic marks when assigning labels, rather using prior annotations. We nonetheless verified after the fact that states with similar labels display similar histone marks, as shown in Figure 2.

We then defined consensus regions of interest, referred to as 'MultiCell' regulatory features. To do so, for each of the labels 'predicted promoter with TSS', 'predicted promoter flank', 'predicted enhancer' and 'CTCF enriched', we computed a summary function, which

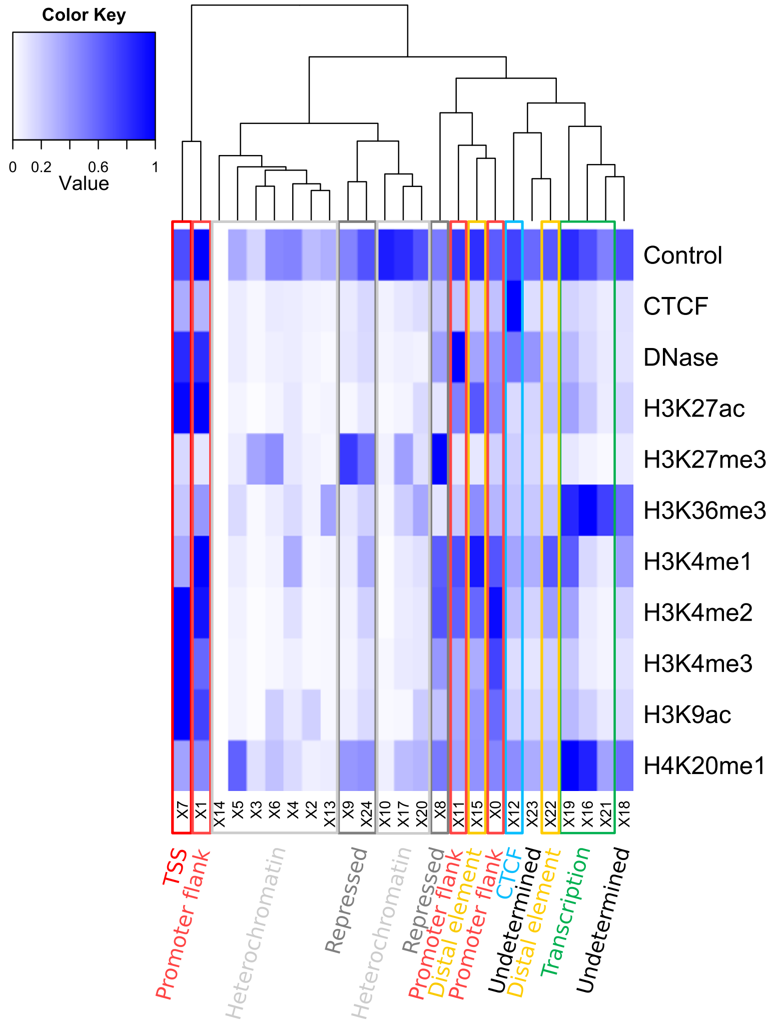

Figure 2 Experimental marks associated with different labels. This heatmap represents the experimental marks and the label associated with each state. The states were defined by Segway, and the labels assigned by the Ensembl Regulatory Build *a posteriori*. Although the label assignment relies mainly on overlaps with known features, the states with the same labels co-cluster based on their experimental marks. The main exception are the promoter flanking states, which cluster either with promoters or with distal *cis*-regulatory elements. In effect, these states tend to represent a mixture of the other two.

represents at any given base pair how many cell types have one of the corresponding segmentation states. We then computed contiguous regions where this summary function is above a threshold, set to optimally fit the global TF binding signal (see Materials and methods section). In addition to these regions, we added regions where TF binding or open chromatin were reported, yet were not covered by the previous annotations.

Finally, the MultiCell features defined above were annotated with cell type-specific activity levels. This activity level was obtained by querying, for each feature, the presence or absence of cell type-specific evidence associated with that feature's label.

We examined the properties of the consensus annotation, as shown in Table 1. The overall coverage of the genome is 12.9%, which is commensurate with previous estimates [25]. The promoters, including attached flanking regions, are by far the largest elements (mean length 4.4 kb), whereas distal enhancers and CTCF binding sites are shorter (respectively 547 and 622 bp on average), but far more numerous (respectively 127,786 and 117,711 elements). Finally, proximal enhancers, defined as flanking regions detached from any promoter, cover the greatest number of bases (160 Mbp in total).

To corroborate our annotation, we compared it with other reference annotations. Of the 217,516 strict TSS calls found with CAGE tags by the FANTOM 5 consortium [31], 88.9% were recovered. Of the 882 validated human VISTA enhancers [32], 92.4% were recalled in our build. Finally, 80.3% of the 38,533 robust enhancers called by FANTOM 5 [33] were covered by one of our annotations.

Discussion

By design, this annotation of the genome is focused on the pragmatic need to define epigenomic markers across samples. Its regulatory features are phenomenological, that is, defined by biochemical signal alone [34]. If only because of the resolution of epigenetic marks (generally at nucleosome scale), they are probably a broad extension of the biochemically active bases in the genome. At the same time, we focused exclusively on the marks associated with transcriptional regulation. This compromise led us to annotating 12.9% of the human genome.

A key parameter that can distort the segmentation is the number of states used by the machine-learning algorithm. Instead of trying to optimize the number of states, we circumvented this issue by focusing on the biologically meaningful labels that are ultimately provided to the user. There are only eight such labels, and Figure 2 illustrates that nearly all labels have more than one underlying state. This suggests that the granularity of the segmentation was sufficient for our purpose, that is, distinguishing these eight labels.

The build process reduces inherently noisy and complex biological data into a tidy and easy to understand summary. Consequently, subtle patterns can be masked from the user. To mitigate this loss of information, all the data used in the Regulatory Build, namely the experimental signal and the segmentations, are available through the Ensembl Browser.

The Ensembl Regulatory Build is by no means a final product, rather a continuing process that will be extended and enriched in the coming years. In future Ensembl releases, we will be importing more and more datasets, covering more cell types, as they are made available. This will provide greater sensitivity to detect transient elements that are only active in a few cell types. Also, we are starting to receive normal cell and tissue data, as opposed to cell lines. Coupled with knowledge of cell differentiation pathways, these data will help illuminate the key epigenomic marks associated with cell fate.

We will also be refining our annotation of regulatory features. The architecture of the Ensembl Regulatory Build process will allow us to take full advantage of ongoing research in machine learning, and genome segmentation in particular. We hope to extend the vocabulary used to describe the elements and the activity levels. For example, we wish to distinguish poised, repressed and closed elements, instead of applying a binary active/inactive notation.

The remaining open question is how to confidently assign gene targets to *cis*-regulatory elements. A number

Table 1 Summary details for the regulatory build in Ensembl release 76

Type	Number	Average length (bp)	Standard deviation (bp)	Total length (Mbp)	Genome coverage (%)
Promoters	16,488	4,369	2,746	72	2.3%
Proximal enhancers	85,526	1,876	1,741	160	5.2%
Distal enhancers	127,786	547	482	70	2.3%
CTCF binding	117,711	622	1,206	73	2.4%
Unannotated transcription factor binding site	27,523	528	628	15	0.5%
Unannotated open chromatin	71,568	502	346	36	1.2%
Total	446,602			399	12.9%

of experimental assays are being investigated, such as statistical correlation [35], chromatin conformation assays [36-38] or expression quantitative trait loci studies [39,40]. The Ensembl framework, which currently holds a consistent relation database of gene transcripts [41], variants [42], and now regulatory elements will be a natural home for this key component of cell biology.

Conclusions

The Ensembl Regulatory Build aims to provide the most up-to-date and comprehensive survey of the regulatory elements of the genome, in the same way the Ensembl Genebuild maintains a reliable summary of known gene sequences. Centralizing datasets from various large-scale projects, we process them with a uniform pipeline, then compute an exhaustive and robust annotation of the regulatory elements of the genome. Although this annotation will likely evolve in the years to come, the regions are already assigned stable identifiers, providing a solid framework for ongoing epigenomic research.

Materials and methods

Source data

We chose to run our segmentation (see below) on a preselected set of ChIP-Seq assays (CTCF, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H4K20me1) along with DNaseI hypersensitivity and a control ChIP-Seq experiment. We therefore downloaded from ENCODE 2 and Epigenomics Roadmap all the raw read datasets produced by ChIP-Seq and DNaseI hypersensitivity experiments on the 18 cell types that had all of the above required assays: A549, DND-41, GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HSMMtube, HUVEC, IMR90, K562, Monocytes-CD14+, NH-A, NHDF-AD, NHEK, NHLF, Osteoblast. Including replicates and control samples, this amounted to 740 datasets, all referenced in the Ensembl homo_sapiens_funcgen_76_38 MySQL database.

Uniform processing of sequencing data

Most studies using epigenomic data present their own analysis and results, which often differ from each other in small, but relevant details. In the current absence of standardized practices, and to make all data as homogeneous as possible, raw sequencing reads from these experiments were processed with a uniform in-house analysis pipeline.

For each ChIP-Seq experiment, the raw sequencing reads were mapped to the GRCh38 human genome assembly using bwa samse [43] with default parameters.

We called punctate peaks using SWEMBL [44]. We filtered SWEMBL peaks on their score, using a fixed permissive threshold (-f 150 -R 0.0005 -d 150), then retained the highest scoring peaks, as defined by the ENCODE Irreproducibility Discovery Rate (IDR) process [45] with an IDR threshold of 0.01 for datasets with more than 100,000, and 0.05 for smaller datasets, as recommended by the IDR developers. To account for large differences in the number of reads between replicates, the number of retained peaks was scaled linearly to half the ratio between the largest and smallest estimated numbers of peaks.

To detect broader regions, such as H3K36me3 and H3K27me3 enrichment, we used CCAT [46]. We filtered out peaks falling within known problematic regions, defined on GRCh38 using the same process as the Duke ENCODE excluded regions [47].

Genome segmentation

The coverage signal was normalized within each dataset using align2rawsignal [48], with options (-w = 180 -n = 5). The segmentation was run across all the resulting datasets using Segway, with options (–num-labels = 25 –num-instances = 10 –resolution = 200 –prior-strength = 1000 –ruler-scale = 200 -m 1,2,3,4,5,6,7,8,9,10,11,12). For performance reasons, training was only computed on the ENCODE pilot regions. The segmentations were masked across the same problematic regions as the peaks.

Computing transcription factor binding densities

For each TF t, we computed a summary function p_t across the genome representing the number of overlapping peak calls at that position divided by the number of assays. This function represents an approximate binomial estimator for the existence of a peak across the observed experiments.

We then computed an overall TF binding probability function assuming approximate independence between the binding probabilities of the different transcription factors:

$$p_{TF} = 1-\prod_{t\in TF}(1-p_t)$$

Assigning labels to segmentation states

For each segmentation state s we constructed a summary function f_s representing for each base pair the number of cell types that are in state s at that position. We computed the enrichment of contiguous regions where f_s was strictly positive for TF binding, TSSs and exonic regions. We also computed the Pearson correlation of f_s to the CTCF density. The state s was then assigned a label using the decision tree represented in Figure 3.

Defining regions of interest through cutoff optimization

We assume that the labels we are interested in, namely *cis*-regulatory elements, promoters and insulators, are correlated to TF binding. Given a cutoff k we computed the enrichment for TF binding signal p_{TF} of regions where f_s was strictly greater than k. If we found a value of k such that this enrichment was greater than 2, then the segmentation state was retained for the next step.

For each label l, we then set a cutoff k_l that maximized the F-score $F_{l,k}$ where S_l is the set of states

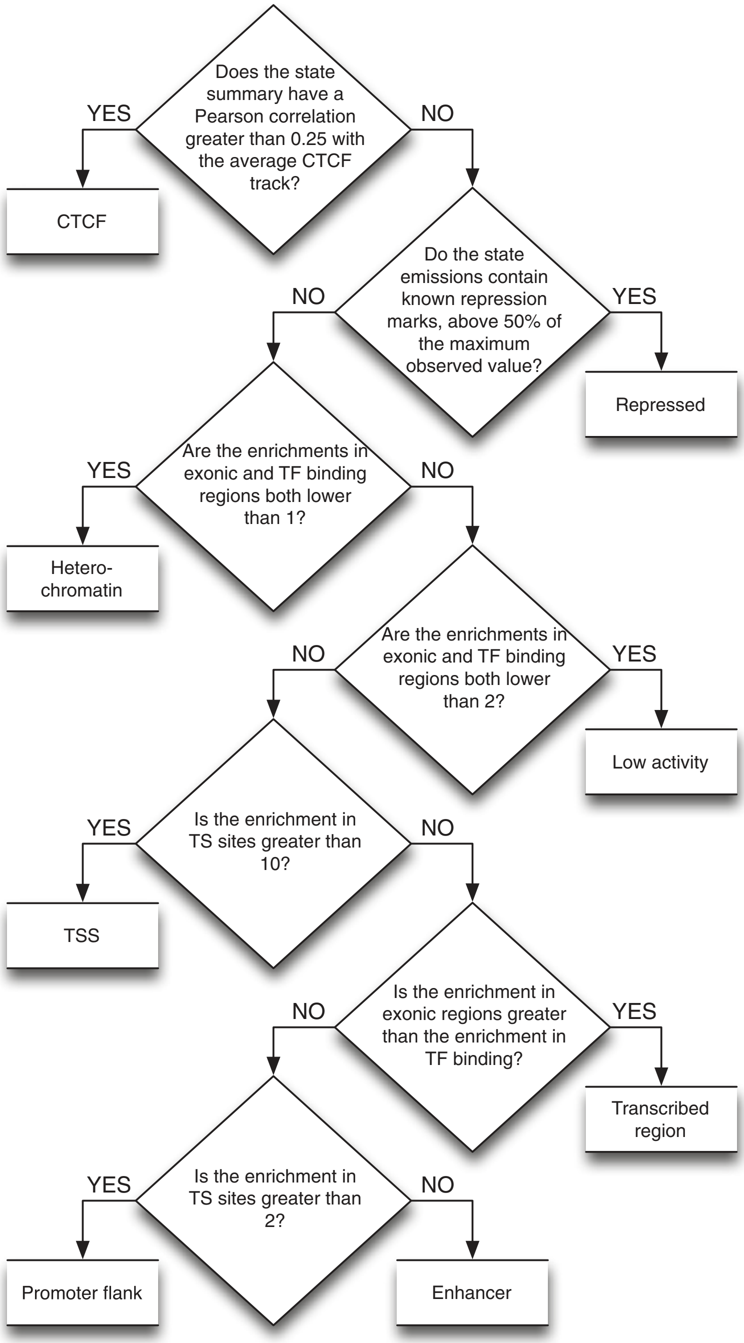
Does the state summary have a Pearson correlation greater than 0.25 with the average CTCF track?
YES
NO
CTCF
Do the state emissions contain known repression marks, above 50% of the maximum observed value?
NO
YES
Repressed
Are the enrichments in exonic and TF binding regions both lower than 1?
YES
NO
Hetero-chromatin
Are the enrichments in exonic and TF binding regions both lower than 2?
NO
YES
Low activity
Is the enrichment in TS sites greater than 10?
YES
NO
TSS
Is the enrichment in exonic regions greater than the enrichment in TF binding?
NO
YES
Transcribed region
Is the enrichment in TS sites greater than 2?
YES
NO
Promoter flank
Enhancer

Figure 3 Decision tree assigning labels to unsupervised segmentation states.

which were assigned that label and passed the above test, and:

$$f_l = \sum_{s \in S_l} f_s$$

$$\delta_{l,k} = \begin{cases} 1 \textit{ if } f_l > k \\ 0 \textit{ otherwise} \end{cases}$$

$$Se = \frac{\int p_{TF}.\delta_{l,k}}{\int p_{TF}}$$

$$Sp = \frac{\int p_{TF}.\delta_{l,k}}{\int \delta_{l,k}}$$

$$F_{l,k} = 2\frac{Se.Sp}{Se + Sp}$$

Having computed k, we computed the contiguous regions where f_l was greater than k_l.

For simplicity, enhancer elements that overlapped promoter flanks were merged into the latter. Promoter-flanking regions that overlapped promoters were merged into the flanks of the promoter element. Because of their structural significance, CTCF binding sites were not merged into overlapping elements.

If any contiguous regions where p_{TF} was greater than 0 did not overlap one of the segmentation-based annotations defined above, it was added into the Build, marked as 'TF binding site'.

Finally, we computed the overlap of all observed open chromatin regions. If one of those did not overlap any of the annotations defined above, it was added into the Build, labeled as 'Open Chromatin'.

Determining cell-specific activity

We then annotated the activity of these features in each cell type with a binary active/inactive label. For each region defined by segmentation data, we searched for an overlap in that cell type's segmentation with a state that had the same label. For each region defined from TF binding sites, we searched for an overlap with a TF binding site detected on that cell type. Finally, for each region defined from open chromatin peaks, we searched for overlap with an open chromatin peak observed in that cell type.

Comparisons

The VISTA enhancers were downloaded from the Ensembl database. The FANTOM5 enhancers and promoters were downloaded from the FANTOM5 servers [49]. These three sets of regions were remapped from GRCh37 to GRCh38 using liftOver [50]. They were then compared with the Ensembl Regulatory Build using bedtools [51].

Software tools

The Ensembl eHive framework [52] was used to maximize the efficient use of available compute resources. All the statistical calculations were performed with the WiggleTools library [53].

Availability and requirements

All Ensembl data and source code are freely available and may be downloaded in their entirety from the Ensembl website [54]. Additionally, the data are available through programmatic Perl, REST interfaces and through the web based Ensembl Biomart. Finally, a track hub [55] contains segmentations, intermediary summary functions and annotations that can be downloaded in bulk. The code used to compute the build is available in script form within the Ensembl Funcgen codebase [56], freely available under an Apache 2 license.

Abbreviations

ChIP: chromatin immuno-precipitation; IDR: Irreproducibility Discovery Rate; TF: transcription factor; TSS: transcription start site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DZ, SW and PF designed the Regulatory Build process, DZ implemented it and SW tested it. NJ and TJ processed the data that were used to create the Ensembl 76 Regulatory Build. DZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Mark Gerstein and Ekta Khurana for constructive suggestions to improve the Build. We thank the other members of the wider Ensembl group at the EBI and the WTSI. Ensembl receives majority funding from the Wellcome Trust (grant number WT095908) with additional funding for specific project components from the National Human Genome Research Institute (1U01 HG004695, 5U01HG004695), the Biotechnology and Biological Sciences Research Council (BB/I025360/2), and the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 282510 (BLUEPRINT). The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 223210 (CISSTEM). This Publication reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein. This work was also supported by the European Commission through grant agreement number 018883 (HEROIC), an Integrated Project funded under the 6th Framework Program; and by the European Commission under grant agreement number 226073 (SLING), an Integrating Activity within Research Infrastructures of the FP7 Capacities Specific Programme. Funding for open access charge: The Wellcome Trust.

Received: 6 October 2014 Accepted: 4 March 2015
Published online: 24 March 2015

References

1. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43:513–8.

2. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature. 2011;471:216–9.
3. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424:147–51.
4. Jaenisch R, Bird A. Epigenetic regulation and gene expression: how the genome integrates integrates intrinsic and environmental signals. Nat Genet. 2003;33:245–54.
5. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genom Human Genet. 2006;7:29–59.
6. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293:1074–80.
7. Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci. 2006;31:89–97.
8. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. 2007;39:457–66.
9. Margueron R, Trojer P, Reinberg D. The key to development: understanding the histone code? Curr Opin Genet Dev. 2005;2:163–76.
10. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol. 2002;1:1–7.
11. Grewal SIS, Jia S. Heterochromatin revisited. Nat Rev Genet. 2007;8:35–46.
12. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447:413–7.
13. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-wide location and function of DNA binding proteins. Science. 2000;290:2306–9.
14. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–37.
15. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448:553–60.
16. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316:1497–502.
17. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, et al. Evidence for an instructive mechanism of *de novo* methylation in cancer cells. Nat Genet. 2006;38:149–53.
18. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
19. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17:877–85.
20. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010. doi: 10.1101/pdb.prot5384.
21. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
22. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
23. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41:827–41.
24. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res. 2007;17:691–707.
25. The ENCODE project consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:54–74.
26. The Roadmap Epigenomics Project. http://www.roadmapepigenomics.org/.
27. The Blueprint Project. http://www.blueprint-epigenome.eu/.
28. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013;41:D48–55.
29. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6.
30. Ernst J, Kellis M. ChromHMM: automating chromatin state discovery and characterization. Nat Methods. 2012;9:215–6.
31. The FANTOM. Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. Nature. 2014;27:462–70.
32. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser - a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35:D88–92.
33. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;27:455–61.
34. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111:6131–8.
35. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
36. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2007;2:988–1002.
37. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
38. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462:58–64.
39. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, et al. A genome-wide association study of global gene expression. Nature. 2007;39:1202–7.
40. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315:848–53.
41. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, et al. The Ensembl automatic gene annotation system. Genome Res. 2004;14:942–50.
42. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, et al. Ensembl variation resources. BMC Genomics. 2010;11:293.
43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
44. SWEMBL. http://www.ebi.ac.uk/~swilder/SWEMBL/.
45. ENCODE IDR Process. https://sites.google.com/site/anshulkundaje/projects/idr.
46. Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, et al. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics. 2010;26:1199–204.
47. The ENCODE Project. http://encodeproject.org.
48. align2rawsignal webpage. https://code.google.com/p/align2rawsignal/.
49. FANTOM5 results. http://enhancer.binf.ku.dk/presets/robust_enhancers.bed and http://enhancer.binf.ku.dk/Pre-defined_tracks.html.
50. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61.
51. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
52. Severin J, Beal K, Vilella AJ, Fitzgerald S, Schuster M, Gordon L, et al. eHive: An Artificial Intelligence workflow system for genomic analysis. BMC Bioinformatics. 2010;11:240.
53. Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek PR. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. Bioinformatics. 2014;30:1008–9.
54. Ensembl website. http://www.ensembl.org.
55. Ensembl Regulatory Build TrackHub. http://ngs.sanger.ac.uk/production/ensembl/regulation/hub.txt.
56. Ensembl Funcgen codebase. http://www.github.com/Ensembl/ensembl-funcgen.